AD-A117 022 SOUTHERN METHODIST UNIV DALLAS TEX DEPT OF STATISTICS F/G 12/1
ROBUST REGRESSION PROCEDURES FOR PREDICTOR VARIABLE OUTLIERS.(U)
MAR 82 D DORSETT, R F GUNST N00014-82-K-0207

UNCLASSIFIED TR-157 NL

END
DATE
FILMED
-82
DTIC

SOUTHERN METHODIST UNIVERSITY

DEPARTMENT OF STATISTICS

DALLAS, TEXAS 75275

ROBUST REGRESSION PROCEDURES FOR
PREDICTOR VARIABLE OUTLIERS

by

Dovalee Dorsett and Richard F. Gunst

Technical Report No. 157
Department of Statistics ONR Contract

March 1982

Research sponsored by the Office of Naval Research
Contract N00014-82-K-0207

DEPARTMENT OF STATISTICS
Southern Methodist University
Dallas, Texas 75275

ROBUST REGRESSION PROCEDURES FOR
PREDICTOR VARIABLE OUTLIERS

Dovalee Dorsett and Richard F. Gunst
Department of Statistics
Southern Methodist University
Dallas, TX 75275

ABSTRACT

Least squares estimators of regression coefficients can be overly sensitive to violations of certain error assumptions; *e.g.*, outliers in the response variable. One solution to the presence of outliers in a data base is to apply univariate robust estimation procedures to the residuals of estimated models. Equally problemmatic as outliers among the response variable are outliers or aberrant values for the predictor variables. Extreme values on individual predictor variables or an unusual combination of predictor variable values for a few observational units can distort least squares estimators even if the error assumptions are valid. This article discusses robust regression procedures, with special emphasis on techniques which are resistant to extreme predictor variable values.

Key Words: M-Estimation, Resistant Estimators, Multicollinearity

1. INTRODUCTION

The adequacy of least squares estimators of regression coefficients is critically dependent on model specification and model assumptions. Although least squares estimators possess powerful theoretical properties (e.g., Seber 1977, Chapter 3) and maintain relative insensitivity to some violations of model assumptions (e.g., Box and Watson 1962), certain model anomalies such as outliers can severely distort least squares estimates (e.g., Gunst and Mason 1980, Section 2.1.3). Robust regression procedures are potentially useful for both detecting and effectively adjusting for outliers.

Outliers among the response or predictor variables can occur for a variety of reasons including transcribing or coding mistakes, unusual experimental conditions, or truly aberrant data values. With large data sets it is often difficult to detect one or a few outliers, particularly if they cluster in the same region of the (p+1)-dimensional space of response and predictor variables. Yet their impact on coefficient estimates can be catastrophic if the outliers lie in strategic corners of the space of response and predictor variables. For these reasons, adaptation of traditional (e.g., maximum likelihood) estimation procedures which could provide protection against outliers are a current focus of research activity.

In this article only Huber's version of M-estimation will be investigated. Other variants of robust regression procedures have been proposed. For example, Andrews (1974) explores M-estimation utilizing a trigonometric weighting function on the residuals. Rupert and Carroll (1980) and Koenker and Bassett (1978) use regression quantiles and trimmed residuals to obtain robust regression estimators. Iman and Conover (1979) adopt rank transforms on the response and the predictor variables in order to reduce the impact of outliers on the prediction of the response variable. Finally, Askin and Montgomery (1980) discuss the combination of robust and biased regression estimators to simultaneously combat the ill effects of outliers and of multicollinearities among the predictor variables.

The sections which follow develop the need for robust regression procedures and suggest methods which can compensate for outliers in the response or the predictor variables. Section 2 of this article outlines robust M-estimation for regression models. Section 3 discusses influence functions and their role in the assessment of robustness properties of estimators. In this section both least squares and M-estimators are shown to be affected by predictor variable outliers. Several proposals for detecting outliers among the predictor variable values and for adjusting regression estimators in order to compensate for these outliers are described in Section 4. Section 5 briefly discusses outlier-induced multicollinearities. A detailed example is given in Section 6 and concluding remarks are made in Section 7.

2. PRELIMINARIES

Write a multiple linear regression model as

$$\underline{Y} = \beta_0 \underline{1} + X\underline{\beta} + \underline{\varepsilon} \ , \qquad (2.1)$$

where $\underline{Y}$ is an n-dimensional vector of observable variables, $\underline{1}$ is a vector of ones, X is a centered ($X'\underline{1} = \underline{0}$) full-column-rank matrix of observations on p nonstochastic predictor variables, β_0 and $\underline{\beta}$ are the unknown constant and p-dimensional vector of regression coefficients, respectively, and $\underline{\varepsilon}$ is an unobservable random error vector. Least squares estimators of the parameters in model (2.1) are obtained by minimizing

$$\sum_{i=1}^{n} \rho(r_i) \ , \qquad (2.2)$$

where $\rho(r_i) = r_i^2$ and $r_i = Y_i - \tilde{\beta}_0 - \underline{u}_i' \tilde{\underline{\beta}}$ is the ith fitted residual based on the estimators $\tilde{\beta}_0$ and $\tilde{\underline{\beta}}$ ($\underline{u}_i'$ is the ith row of X). Since $\rho(\cdot)$ is differentiable one can easily show that minimization of (2.2) is equivalent to solving the following system of (p+1) homogeneous equations (the "normal equations"):

$$\sum_{i=1}^{n} \psi(r_i) = 0 \ , \quad \sum_{i=1}^{n} X_{ij}\psi(r_i) = 0 \qquad j = 1,2,\ldots,p \qquad (2.3)$$

where $\psi(t) = d\rho(t)/dt \propto t$. The resulting least squares estimators are

$$\hat{\beta}_0 = \bar{Y} \quad \text{and} \quad \underline{\hat{\beta}} = (X'X)^{-1}X'\underline{Y}. \tag{2.4}$$

If $\varepsilon_i \sim NID(0,\sigma^2)$, *the least squares estimators are maximum like*lihood estimators since $\rho(\varepsilon) = -2\sigma^2 \ln[f_\sigma(\varepsilon)] + c$, where $f_\sigma(\varepsilon)$ is the density function for a $N(0,\sigma^2)$ variate and c is a constant which does not depend on β_0 and $\underline{\beta}$

Robust M-estimators seek to reduce the influence of aberrant response values while retaining an equivalence with maximum likelihood estimators when no such wild response values occur. This is accomplished by selecting a function $\rho(\cdot)$ which will leave "typical" residuals unchanged but will lessen the influence of large residuals on the solution of eqns. (2.3). Most M-estimation procedures require that $\rho(\cdot)$ be convex, nonmonotone, and that it possess a bounded, continuous derivative $\psi(\cdot)$. The convexity and monotonicity properties are imposed to insure unique solutions while the boundedness and continuity of $\psi(\cdot)$ insure that the estimator cannot be dominated by an extremely large residual (boundedness) and that small changes in residuals cannot produce large changes in the resulting estimates (continuity). Existence of higher-order derivatives of $\rho(\cdot)$ are desirable for theoretical derivations of asymptotic properties of M-estimators.

Huber (1964) popularized the use of a robust M-estimator which can be defined in terms of the following function $\rho(\cdot)$:

$$\rho(r_i) = \begin{cases} \frac{1}{2} r_i^2 & |r_i| \leq c \\ c|r_i| - \frac{1}{2} c^2 & |r_i| > c \end{cases} \tag{2.5}$$

Equivalently, the estimator can be defined as the solution of eqns. (2.3) when the following $\psi(\cdot)$-function is used

$$\psi(r_i) = \begin{cases} r_i & |r_i| \leq c \\ c\cdot \text{sign}(r_i) & |r_i| > c \end{cases} \tag{2.6}$$

The value of c in eqn. (2.6) is often chosen to be a multiple of a robust estimator of σ. Note that $\psi(\cdot)$ is bounded by $\pm c$ and that if all the residuals are less than c in magnitude the solution of eqn. (2.3) using this $\psi(\cdot)$-function will be identical with the least squares (maximum likelihood) estimator.

Computationally, several decisions must be reached before M-estimates can be obtained. First, initial estimates of β_0 and $\underline{\beta}$ must be determined so residuals can be calculated and inserted into eqns. (2.3). Second a choice for c and perhaps a robust estimator of σ must be selected for use with $\psi(\cdot)$. Finally, a computational scheme for iterating to obtain new estimates must be devised. These considerations are discussed in Dutter (1975, 1977) and Huber (1981, Section 7.8) and will not be explored in detail here; however, we will briefly outline one adaption of their computational scheme.

In order to insure convergence, that a minimum is reached, and to allow for the simultaneous estimation of β_0, $\underline{\beta}$ and σ^2, Dutter (1975, 1977) and Huber (1981) elect not to minimize eqn. (2.2) but instead choose to minimize

$$n^{-1}\sum_{i=1}^{n}[\rho(r_i/\tilde{\sigma}) + a]\tilde{\sigma} \quad , \tag{2.7}$$

or, equivalently, to solve the following system of (p+2) equations

$$\sum_{i=1}^{n}\psi(r_i/\tilde{\sigma}) = 0 \quad , \quad \sum_{i=1}^{n}X_{ij}\psi(r_i/\tilde{\sigma}) = 0 \quad j = 1,2,\ldots,p \tag{2.8}$$

and

$$n^{-1}\sum_{i=1}^{n}\chi(r_i/\tilde{\sigma}) = a \quad , \tag{2.9}$$

where $\chi(t) = t\psi(t) - \rho(t)$. Equations (2.3) and (2.8) are identical if r_i is replaced in the former set by $r_i/\tilde{\sigma}$. Since the residuals are standardized in eqns. (2.8) by an estimate of scale, the value of c in eqn. (2.6) need not depend on $\tilde{\sigma}$ and is often chosen to be 1.5. The value of a is selected so eqn. (2.9) will yield a consistent estimator of σ when $\varepsilon \sim N(0,\sigma^2)$; viz.; $a = (n-p-1)E[\chi[\varepsilon/\sigma)]/n$.

Iterating with eqns. (2.8) and (2.9) is relatively straightforward. Let $\tilde{\underline{\theta}}_{(k)}$ denote the estimates of $\underline{\theta}' = (\beta_0, \underline{\beta}')$ obtained on the kth iterate and let $\tilde{\sigma}^2_{(k)}$ denote the corresponding estimate of σ^2. From eqn. (2.9), a new estimate of σ^2 is

$$\tilde{\sigma}^2{}_{(k+1)} = (na)^{-1}\sum_{i=1}^{n}\chi(r_i/\tilde{\sigma}_{(k)})\tilde{\sigma}^2_{(k)} \quad , \tag{2.10}$$

where for ease of notation we let r_i denote the ith residual obtained from the kth iteration. By letting $\phi(t) = \psi(t)/t$, eqns. (2.8) can be rewritten as

$$\sum_{i=1}^{n} \phi(r_i/\tilde{\sigma}) \cdot (r_i/\tilde{\sigma}) = 0 \quad , \quad \sum_{i=1}^{n} X_{ij} \; \phi(r_i/\tilde{\sigma}) \cdot (r_i/\tilde{\sigma}) = 0 \qquad (2.11)$$

or as

$$\hat{\underline{\theta}} = (Z'\Phi Z)^{-1} Z'\Phi \underline{Y} \quad , \qquad (2.12)$$

where $Z = [\underline{1}, X]$ and $\Phi = \mathrm{diag}(\phi(r_1/\tilde{\sigma}), \ldots, \phi(r_n/\tilde{\sigma}))$. Equation (2.12) is simply a weighted least squares estimator of $\underline{\theta}$ in which the stochastic weights are $\phi(r_i/\tilde{\sigma})$. Using the residuals from the kth iteration and $\tilde{\sigma}^2_{(k+1)}$ from eqn. (2.10), $\tilde{\underline{\theta}}_{(k+1)}$ is found from this weighted least squares estimator.

Based on the foregoing, iterative estimation of the parameters of model (2.1) can be based on the following sequence of steps:

1. Obtain initial estimates of β_0 and $\underline{\beta}$ from eqns. (2.4) or from one of the estimators proposed in Section 4,
2. Use either the least squares estimate of σ or some robust estimate of scale; e.g., $\tilde{\sigma} = \{\mathrm{median}|r_i^*|\}/.6745$, where $r_i^* = r_i - \mathrm{median}\{r_i\}$, (Andrews, et al. 1972),
3. Calculate $\tilde{\sigma}^2_{(k+1)}$ from eqn. (2.10) using $c = 1.5$,
4. Update the estimates of β_0 and $\underline{\beta}$ with the weighted least squares estimator (2.12),
5. Repeat steps 3 and 4 until satisfactory convergence is reached.

This algorithm for finding robust regression estimates provides good protection against aberrant response (error) terms. Reasons for this protection, apart from the informal discussions given above, can be readily appreciated by examining the influence functions corresponding to least squares and M-estimators. At the same time, the lack of "resistance" of both of these estimators to outliers in the predictor variables can be seen from the influence functions. We now turn to this more formal evaluation of the sensitivity of regression estimators to violations of model assumptions.

3. INFLUENCE FUNCTIONS

Hampel (1968, 1974) introduced the use of influence functions for studying robustness properties of estimators. The local behavior of an estimator in a neighborhood of the assumed underlying distribution is studied by first expressing the estimator as a functional on a space of probability distributions. Then the influence function of the estimator is defined to be the derivative of the functional evaluated at the assumed distribution. Not only can idealized or "parametric" influence functions be defined in this manner but empirical influence functions can also be defined in terms of empirical distribution functions. Before turning to regression models, these concepts will be illustrated on a simple location model.

Let $T(F)$ denote a real-valued functional defined on a subset of probability distributions, $F \epsilon \mathcal{F}$. For example, the mean functional can be defined as

$$\int (x-T(F))dF(x) = 0 \quad , \tag{3.1}$$

yielding $T(F) = \mu = \int x \, dF(x)$. If F_n is an empirical c.d.f. based on a random sample of size n from F, an estimator of T(F) can be derived from eqn. (3.1) as

$$\int (x-T(F_n))dF_n(x) = 0 \tag{3.2}$$

or $T(F_n) = \hat{\mu} = n^{-1}\sum x_i$. The functional T(F) can be viewed either as a parametric analogue to the finite-sample estimator (3.2) or as a limiting estimator for very large sample sizes.

Consider next the effect of an outlier, x_0, on T(F) and $T(F_n)$. In the space of probability distributions an outlier can be modeled as a mixture distribution

$$F^{\alpha}(x) = (1-\alpha)F(x) + \alpha H_0(x), \qquad 0 \leq \alpha \leq 1 \tag{3.3}$$

where

$$H_0(x) = \int_{-\infty}^{x} \delta_0(t)dt$$

and $\delta_0(t)$ is a probability density function for the contaminant. For the remainder of this section we will assume that $\delta_0(t)$ assigns point mass to x_0. Using this contaminated (point mass) distribution function the influence of x_0 on the estimator can be assessed.

A measure of the impact of an outlier x_0 on the estimator

$\dot{T}(F)$ is the "influence function" which is defined to be

$$\dot{T}(F) = \lim_{\alpha \to 0^+} \frac{T(F^\alpha)-T(F)}{\alpha} , \tag{3.4}$$

where $\dot{T}(F)$ can be viewed as a directional derivative of $T(F)$ in the direction of x_0 if the limit exists and is unique as the limit is taken from positive and negative directions. An empirical influence function can be defined in a similar fashion simply by replacing F with F_n in eqn. (3.4).

Contamination of the assumed distribution by the outlier x_0 distorts the estimator. For large samples $T(F^\alpha) = \mu + \alpha(x_0-\mu)$ and $\dot{T}(F) = x_0-\mu$. Thus the influence (distortion) of the estimator is proportional to $x_0-\mu$. For finite samples, $T(F_n^\alpha) = \bar{x}_n + \alpha(x_0-\bar{x}_n)$ and $\dot{T}(F_n) = x_0-\bar{x}_n$, where $\bar{x}_n = n^{-1}\sum x_i$. Note that in either case the influence functions are unbounded functions of the contaminant x_0; consequently, a single gross outlier can have a devasting effect on the estimator even if the outlier occurs with relatively small likelihood (α).

These results contrast with robust M-estimation in that the latter estimators possess bounded influence functions and thereby limit the distortion an outlier can cause. Robust M-estimator functionals in the location model satisfy the equation

$$\int \psi(x-T(F))dF(x) = 0 , \tag{3.5}$$

which reduces to eqn. (3.1) when $\psi(t) = t$. Replacing $F(x)$ by $F^\alpha(x)$, differentiating eqn. (3.5) implicitly, and evaluating the derivative at $\alpha = 0$ yields

$$\dot{T}(F) = \frac{\psi(x_0 - T(F))}{\int \dot{\psi}(x - T(F))\,dF(x)}, \tag{3.6}$$

where $\dot{\psi}(t) = d\psi(t)/dt$. The influence function (3.6) is proportional to $\psi(\cdot)$ and is thereby a bounded function of x_0. Analogous properties hold for the empirical influence function.

Turning now to the regression model (2.1), the regression functional can be written as

$$\int Z'(\underline{Y} - X\underline{T}(F))\,dF(\underline{Y}) = \underline{0}, \tag{3.7}$$

where $Z = [\underline{1}, X]$ and $F(\underline{Y})$ represents the c.d.f. of a multivariate normal density function, $\underline{Y} \sim N(Z\underline{\theta}, \sigma^2 I)$ with $\underline{\theta}' = (\beta_0, \underline{\beta}')$. This functional can be rewritten as

$$\underline{T}(F) = \underline{\theta} = (Z'Z)^{-1} Z' \int \underline{Y}\, dF(\underline{Y}). \tag{3.8}$$

It is important to realize that the response vector $\underline{Y}$ represents a single observation from this multivariate normal distribution and not n independent observations from a univariate distribution. Thus an appropriate contaminated distribution for this functional is

$$F^{\alpha}(\underline{Y}) = (1-\alpha)F(\underline{Y}) + \alpha H_0(\underline{Y}) \tag{3.9}$$

where $H_0(\underline{Y})$ is a c.d.f. for the contaminated distribution of an n-dimensional outlier $\underline{Y}_0 = Z\underline{\theta} + \underline{\varepsilon}_0$. The error $\underline{\varepsilon}_0$ does not follow the assumed $N(\underline{0}, \sigma^2 I)$ distribution and could be partially or completely deterministic. Single response outliers can be modeled by defining (n-1) of the elements of $\underline{\varepsilon}_0$ to have the assumed $NID(0, \sigma^2)$ error distribution and the remaining one to have a different distribution (perhaps deterministic). The influence function corresponding to eqns. (3.7) and (3.9) is

$$\dot{\underline{T}}(F) = (Z'Z)^{-1}Z'(\underline{Y}_0-Z\underline{\theta})$$

$$= (Z'Z)^{-1}Z'\underline{\varepsilon}_0 \ . \qquad (3.10)$$

This influence function reveals that the distortion in the functional (3.8) is proportional to the error vector, $\underline{\varepsilon}_0 = \underline{Y}_0 - Z\underline{\theta}$, and is an unbounded function of the elements of the contaminant $\underline{Y}_0$. Thus, as in the location model, regression estimators can be severely distorted by gross outliers.

Corresponding to the functional (3.8), a regression functional for a robust M-estimator is

$$\int Z'\underline{\Psi}(\underline{Y}-Z\underline{T}(F))dF(\underline{Y}) = \underline{0} \ , \qquad (3.11)$$

where $\underline{\Psi}(\underline{t}) = (\psi(t_1),\ldots,\psi(t_n))'$ for some robust $\psi(\cdot)$-function. Using eqn. (3.9) as the contaminated distribution produces the following expression for the influence function $\dot{\underline{T}}(F)$:

$$Z'\int\dot{\Psi}(\underline{Y}-Z\underline{\theta})dF(\underline{Y})Z\dot{\underline{T}}(F) = Z'\underline{\Psi}(\underline{Y}_0-Z\underline{\theta}) \ , \qquad (3.12)$$

where $\dot{\Psi}(\underline{t}) = \mathrm{diag}(\dot{\psi}(t_1),\ldots,\dot{\psi}(t_n))$. As with eqn. (3.6) for the location model, the influence function in eqn. (3.12) is proportional to $\underline{\Psi}(\underline{Y}_0-Z\underline{\theta})$ and is therefore a bounded function of the elements of $\underline{Y}_0$. A similar derivation for the empirical influence function $F_1(\underline{Y})$ yields

$$Z'\dot{\Psi}(\underline{Y}-Z\underline{\tilde{\theta}})Z\dot{\underline{T}}(F_1(\underline{Y})) = Z'\underline{\Psi}(\underline{Y}_0-Z\underline{\tilde{\theta}})$$

or

$$\underline{\dot{T}}(F_1(\underline{Y})) = [Z'\dot{\Psi}(\underline{Y}-Z\underline{\tilde{\theta}})Z]^{-1}Z'\underline{\Psi}(\underline{Y}_0-Z\underline{\tilde{\theta}}) \ , \qquad (3.13)$$

where $\underline{\tilde{\theta}}$ is the robust M-estimator (2.11).

Hampel (1974) and Huber (1981) justify the need for bounded influence functions. Intuitively, the above derivations show that estimators can be severely distorted by gross contaminants even if their likelihood of occurrence is small. Robust M-estimators bound the influence functions and thereby limit the change which an errant point can produce in an estimator. Of special importance is the safeguard that robust M-estimators provide against catastrophic distortions by outliers in the response variable for either location or regression models.

Although robust M-estimators provide protection from contaminated distributions for the response variable, it should be apparent from eqns. (3.12) and (3.13) that no specific protection is offered for aberrant predictor variable values. That outliers in the predictor variables is as insidious a problem as outliers in the response variable can be illustrated with a simple example. Supppose eqn. (2.1) represents a single-variable, no-intercept model: $Y_i = \beta X_i + \varepsilon_i$. Rewrite eqn. (2.3) as

$$\psi(Y_1-X_1\tilde{\beta}) + X_1^{-1}\sum_{i=1}^{n} X_i\psi(Y_i-X_i\tilde{\beta}) = 0.$$

If the response variables are held fixed and $X_1 \to \infty$, the second term of this equation is driven to zero. Consequently, eqn. (2.3) reduces to

$$\psi(Y_1 - X_1\tilde{\beta}) = 0$$

which has a solution $\tilde{\beta} = X_1^{-1}Y_1 \to 0$. Thus regardless of the value of β, the estimator of β approaches 0 for both least squares, $\psi(t) = t$, and for any robust M-estimator possessing the properties described in Section 2; in particular, nonmonotonic continuous $\psi(\cdot)$-functions for which $\psi(0) = 0$, including eqn. (2.6).

This example illustrates that a single errant predictor variable value can have as catastrophic an effect on estimation of parameters for regression models as can outliers in the response variable. The next section examines several proposals for dealing with aberrant predictor variable values.

4. PROPOSED SOLUTIONS

A natural solution to the problem of outliers in the predictor variables is to weight each predictor variable in a fashion similar to M-estimation on the response variable. Accordingly, one could replace X_{ij} by $\psi^j(X_{ij})$, where

$$\psi^j(X_{ij}) = \begin{cases} X_{ij} & |X_{ij}| \le c_j \\ c_j \cdot \text{sign}(X_{ij}) & |X_{ij}| > c_j \end{cases} \qquad (4.1)$$

$c_j = 1.5s_j$, and s_j is a robust measure of scale for the n observations on X_j. One could also center X_j with robust estimate of location prior to forming the $\psi^j(\cdot)$.

Another proposal (Mallows 1973, Denby and Larson 1977) is to replace X_{ij} by $\psi_M(X_{ij})$, where

$$\psi_M(X_{ij}) = X_{ij}\prod_{k=1}^{p} q_k(X_{ik}) \tag{4.2}$$

and

$$q_k(X_{ik}) = \begin{cases} (X_{Q_1}-X_{Q_2})/(2X_{ik}-X_{Q_1}-X_{Q_2}) & r^*_{ik} < Q_1 \\ 1 & Q_1 \le r^*_{ik} \le Q_2 \\ (X_{Q_2}-X_{Q_1})/(2X_{ik}-X_{Q_1}-X_{Q_2}) & r^*_{ik} > Q_2 \end{cases}$$

In this weighting scheme $Q_1 = 1 + [.06n]$, $Q_2 = n+1-Q_1$, $[t]$ = greatest integer $\le t$, and r^*_{ik} is the rank of X_{ik} and X_{Qj} is the jth sample percentile values of X_k.

Each of the $\psi(\cdot)$-functions (4.1) and (4.2) should be effective protection against a few outliers in individual predictor variables. Neither of these weighting schemes might be effective if the outliers are due to rows of X lying in an extreme corner of the observed predictor variable space but not having an extreme value on any individual predictor variable. Denby and Larson (1977) observed that the estimators in their simulation did not perform satisfactorily when a single outlier among the predictor variables was induced by adding a large quantity to each of two predictor variables. Both M-estimation and Mallows adaptation (4.2), among others, were unable to successfully account for the effect of the two-dimensional outlier. We now propose two additional alternatives for multidimensional outliers.

Detection of extreme predictor variable values, or combinations of predictor vairables, is the first step in rectifying the estimation problems they produce. Abnormally large or small values for individual predictor variables are relatively easy to detect from (perhaps robust) summary statistics. For example, some computer programs automatically "flag" observations which are further than two or three standard deviations from the mean. Examination of the weights $\psi^j(X_{ij})$ or $q_k(X_{ik})$ are also useful for detecting outliers in one dimension. For outliers in two or more dimensions other techniques are needed.

Hoaglin and Welsch (1978) popularized the use of a matrix referred to as the "hat matrix" to detect outliers among the predictor variables. The hat matrix is so named because it transforms the response vector into the least squares prediction vector $\hat{\underline{Y}} = H\underline{Y}$ where

$$\begin{aligned} H &= Z(Z'Z)^{-1}Z' \\ &= n^{-1}\underline{1}\,\underline{1}' + X(X'X)^{-1}X' \quad . \end{aligned} \tag{4.3}$$

Diagonal elements of the hat matrix are

$$h_{ii} = n^{-1} + \underline{u}_i'(X'X)^{-1}\underline{u}_i \tag{4.4}$$

where the quadratic form in $\underline{u}_i$ represents a (squared) Mahalanobis distance of the ith row of X from the centroid of the predictor variable space. Large values of h_{ii} indicate rows of X which lie in extreme regions of the observed predictor variable space. Anomalous values on one or more predictor variables can be detected by the h_{ii}. Since the predictor equation for the ith response variable can be written as

$$\hat{Y}_i = h_{ii}Y_i + \sum_{j\neq i} h_{ij} Y_j , \qquad (4.5)$$

the h_{ii} are a direct measure of the relative importance of Y_i in predicting its own value. Due to the importance of the diagonal elements of H in detecting multidimensional outliers and assessing the influence of Y_i on $\hat{Y}_i$, they have been termed "leverage values."

The hat matrix H is idempotent; consequently, the leverage values are constrained to the interval [0,1]. The more extreme a row of X is relative to the other rows of X, the closer the corresponding leverage value is to 1. For example, if model (2.1) contains a single predictor variable

$$h_{ii} = n^{-1} + (X_i-\bar{X})^2 / \sum_{k=1}^{n} (X_k-\bar{X})^2 .$$

Observe that if $X_i = \bar{X}$, $h_{ii} = n^{-1}$ but if X_1 is very large in magnitude

$$h_{ii} = n^{-1} + \frac{(1 - X_i^{-1}\bar{X})^2}{(1 - X_i^{-1}\bar{X})^2 + \sum_{j\neq i} (X_i^{-1}X_j - X_i^{-1}\bar{X})^2}$$

$$\simeq n^{-1} + \frac{(1 - n^{-1})^2}{(1 - n^{-1})^2 + (n - 1)(-n^{-1})^2} = 1 .$$

In the previous section it was shown that as $X_1 \to \infty$ the least squares estimator $\hat{\beta}$ approaches zero. From eqn. (4.5) or by directly evaluating $\hat{Y}_1 = X_1\hat{\beta}$ one can show that $\hat{Y}_1 \to Y_1$ as $X_1 \to \infty$. In general, if $\underline{u}_1'$ is a single outlier in X the corresponding predicted response will be almost uniquely determined by, and equal to, its observed response. Concommitant with near perfect prediction of Y_i when $h_{ii} \approx 1$ will often occur severe distortion of one or more of the coefficient estimates.

Multivariate outliers are detectable not only by large leverage values but also in the normalized principal components of X. Let X_S denote the standardized ($X_S'X_S$ is in correlation form) matrix of predictor variables. Further, let $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_p$ denote the latent roots of $X_S'X_S$ and $\underline{V}_1, \underline{V}_2, \ldots, \underline{V}_p$ the corresponding latent vectors. The jth normalized principle component of X is $\underline{m}_j = \ell_j^{-\frac{1}{2}} X_S\underline{V}_j$. An extreme row of X causes an elongation of one of the component axes and a large element in the corresponding normalized principle component. Since the component vectors are mutually orthonormal, univariate weights such as eqn. (4.1) or (4.2) could prove effective in obtaining estimators of the principal component coefficients $\gamma_j = \ell_j^{\frac{1}{2}}\underline{V}_j'\underline{\beta}$ which are resistant to outliers in the predictor variables. Inverse transformations could then produce resistant estimators of the β_j.

Another alternative to the above proposals is a direct weighting of the rows of X. Consider weighting the kth row of X by a factor $\omega_k^{\frac{1}{2}}$, where $0 \leq \omega_k \leq 1$. Then model (2.1) is replaced by

$$\underline{Y} = \beta_0^* \underline{1} + X_\Omega \underline{\beta} + \underline{\varepsilon} , \tag{4.6}$$

where $\beta_0^* = \beta_0 + n^{-1} \underline{1}'\Omega^{\frac{1}{2}}X\underline{\beta}$, $\Omega = \text{diag}(1,\ldots,1,\omega_k,1,\ldots 1)$, and $X_\Omega = (I - n^{-1}\underline{1}\,\underline{1}')\,\Omega^{\frac{1}{2}}X$ (i.e., X_Ω contains centered values of the matrix $\Omega^{\frac{1}{2}}X$). Underlying the use of model (4.6) is the assumption that when inordinately large predictor variable values occur this model is more reasonable an expression of the relationship between response and predictor variables than is model (2.1). The impact of this model on the estimation of parameters is that eqns. (2.8) become

$$\sum_{i=1}^{n} \psi(r_i/\tilde{\sigma}) = 0 \ , \ \sum_{i=1}^{n} \psi_\Omega(X_{ij})\psi(r_i/\tilde{\sigma}) = 0 \quad j = 1,2,\ldots,p) \tag{4.7}$$

where

$$\psi_\Omega(X_{ij}) = \omega_i^{\frac{1}{2}} X_{ij} - \bar{X}_{j\Omega} \quad , \quad \bar{X}_{j\Omega} = n^{-1} \sum_{i=1}^{n} \omega_i^{\frac{1}{2}} X_{ij}$$

and $\omega_i = 1$, $i \neq k$. Iterative solutions of eqns. (4.7) and (2.9) can be obtained as in Section 2. Note that $\tilde{\beta}_0$ can be obtained from $\underline{\tilde{\beta}}_0^*$ and $\underline{\tilde{\beta}}$ by the following relationship

$$\begin{aligned}\tilde{\beta}_0 &= \tilde{\beta}_0^* - n^{-1} \sum_{j=1}^{p} \bar{X}_{j\Omega}\tilde{\beta}_j \\ &= \tilde{\beta}_0^* + n^{-1}(1 - \omega_k^{\frac{1}{2}}) \sum_{j=1}^{p} X_{kj}\tilde{\beta}_j\end{aligned}$$

and that asymptotically $\tilde{\beta}_0$ and $\tilde{\beta}_0^*$ are identical.

Leverage values for the weighted model (4.6) can be calculated from

$$H_\Omega = n^{-1}\underline{1}\,\underline{1}' + X_\Omega(X_\Omega'X_\Omega)^{-1}X_\Omega' \ .$$

As a function of ω_k and the original leverage values, the kth leverage value of H_Ω is

$$h_{kk}(\omega_k) = n^{-1} + (h_{kk} - n^{-1})t_2^2/[1 - t_1(h_{kk} - n^{-1})] \qquad (4.8)$$

where $t_1 = (1-\omega_k) + n^{-1}(1-\omega_k^{1/2})^2$ and $t_2 = n^{-1}(1+(n-1)\omega_k^{1/2})$. For large sample sizes $t_1 \simeq 1-\omega_k$ and $t_2 \simeq \omega_k^{1/2}$, yielding an approximation to eqn. (4.8):

$$h_{kk}(\omega_k) \simeq \frac{\omega_k h_{kk}}{1-(1-\omega_k)h_{kk}} . \qquad (4.9)$$

Note that for large sample sizes and $\omega_k < 1$, $h_{kk}(\omega_k) < h_{kk}$; moreover, algebraic manipulation of eqn. (4.8) allows one to verify that the same property holds for all sample sizes.

Equation (4.9) provides a rationale for selecting a value of ω_k. Suppose one wishes to fix the leverage value of the kth row of X_Ω to be a suitably small or moderate value η, $n^{-1} \le \eta \le h_{kk}$. By setting $h_{kk}(\omega_k) = \eta$ in eqn. (4.9) one can solve for a value of ω_k:

$$\omega_k = \frac{\eta(1-h_{kk})}{h_{kk}(1-\eta)} . \qquad (4.10)$$

Note that setting $\eta = n^{-1}$ yields $\omega_k \simeq 0$ for moderate to large sample sizes ; i.e., setting $\eta = n^{-1}$ in eqn. (4.10) results in replacement of $\underline{u}_k'$ by $\underline{0}'$ (approximately). Similarly, setting $\eta = h_{kk}$ yields $\omega_k = 1$; i.e., $\underline{u}_k'$ is left unchanged in X.

To illustrate the effect of this type of weighting scheme, let us return to the single-variable, no intercept model and assume that X_1 is an outlier. For this model,

$$h_{ii} = X_i^2 / \sum_{k=1}^{n} X_k^2 \quad , \quad h_{11}(\omega_1) = \frac{\omega_1 h_{11}}{1 - (1-\omega_1)h_{11}}$$

and

$$\omega_1 = \frac{\eta(1-h_{11})}{h_{11}(1-\eta)}$$

for some specified value η of $h_{11}(\omega_1)$. Then

$$\psi_\Omega(X_1) = \left(\frac{\eta}{1-\eta}\right)^{\frac{1}{2}} \left(\sum_{i\neq 1}^{n} X_i^2\right)^{\frac{1}{2}}$$

and $\psi_\Omega(X_i) = X_i$ for $i \neq 1$. Thus if $\psi(t) = t$ in eqn. (4.7) the least squares estimator for the weighted model (4.6) is

$$\tilde{\beta} = \sum_{i=1}^{n} \psi_\Omega(X_i) Y_i / \sum_{i=1}^{n} \psi_\Omega(X_i)^2$$

$$= \frac{[\eta \sum_{i\neq 1}^{n} X_i^2/(1-\eta)]^{\frac{1}{2}} Y_1 + \sum_{i\neq 1}^{n} X_i Y_i}{(1-\eta)^{-1} \sum_{i\neq 1}^{n} X_i^2}$$

which yields a finite and (generally) nonzero estimate of β, unlike the solutions given in the last section. Iterative solution of eqn. (4.7) for $\psi(t)$ given by eqn. (2.4) will likewise not degenerate to a zero solution as $X_1 \to \infty$.

Generalizations of this procedure to two or more outliers in X are possible and can follow a development similar to the above. The theoretical results are far more complex and iterative schemes need to be developed in order to solve for weights which will enable two or more leverage values to be simultaneously satisfied. Although cruder and only an approximation, a simpler approach to situations in which two or more outliers are present would be to use eqn. (4.10) as a guide to an initial specification of weights and then alter the weights jointly until a satisfactory combination of leverage values is attained.

5. OUTLIER-INDUCED MULTICOLLINEARITIES

Observations which possess very large values on two or more predictor variables can induce multicollinearities among the predictor variables. Unlike the usual situation in which all observations conform to the multicollinearity, an outlier-induced multicollinearity is an artifice of the outliers and not a true indication of a redundancy among the predictor variables. Deletion of the outliers from the data base destroys this type of multicollinearity.

The effects of an outlier-induced multicollinearity on a regression analysis are similar to those resulting from a true multicollinearity. Coefficient estimates tend to be too large in magnitude, their signs tend to be determined by the multicollinearity itself and not the true relationship between response and preditdor variables, and the variances of coefficient estimators for multicollinear predictor variables can be orders of magnitude larger than if the predictor variables were not multicollinear. For example, if eqn. (2.1) represents a two-variable, no-intercept model the estimating equations (2.3) become

$$\sum_{i=1}^{n} X_{ij}\psi(Y_i-\tilde{\beta}_1 X_{i1}-\tilde{\beta}_2 X_{i2}) = 0 \qquad j=1,2. \tag{5.1}$$

If one now lets $X_{1j} \to \infty$, $j=1,2$, and restricts X_{11} and X_{12} so $X_{11}/X_{12} = 1$, eqns. (5.1) reduce to

$$\psi(Y_1-(\tilde{\beta}_1+\tilde{\beta}_2)X_{11}) = 0 \ , \tag{5.2}$$

which has as solutions $\tilde{\beta}_1+\tilde{\beta}_2 \approx Y_1/X_{11} \to 0$. The limiting solution of eqn. (5.2) forces $\tilde{\beta}_1 \approx -\tilde{\beta}_2$ but $\tilde{\beta}_1$ and $\tilde{\beta}_2$ can have almost any magnitude, regardless of the true values of β_1 and β_2. This type of ambiguous solution is characteristic of least squares estimation when predictor variables are multicollinear. Since eqns. (5.1) and (5.2) are also estimating equations for M-estimators, robust M-estimation can also suffer ill-effects of outlier-induced multicollinearities.

Biased estimation is frequently offered as a solution to estimation with multicollinear predictor variables. An attractive alternative to biased estimation when multicollinearities are caused by a few outliers is robust estimation; however, it should be apparent from the above example that the robust procedures must be resistant to predictor variable outliers.

As with the single-variable example in the previous section, weighting a single outlier provides protection against the domination of the estimator by the outlier. For this two-variable example the weighted M-estimator is obtained from the equations

$$\sum_{i=1}^{n} \psi_\Omega(X_{ij})\psi(Y_i - \tilde{\beta}_1\psi_\Omega(X_{i1}) - \tilde{\beta}_2\psi_\Omega(X_{i2})) = 0 \qquad j=1,2 \quad ,$$

where $\psi_\Omega(X_{ij}) = X_{ij}$ for $i \neq 1$ and $\psi_\Omega(X_{1j}) = \omega_1^{\frac{1}{2}} X_{1j}$. If $X_{1j} \to \infty$ with $X_{11}/X_{12} = 1$,

$$\psi_\Omega(X_{1j}) = \omega_1^{\frac{1}{2}} X_{1j} \qquad j=1,2$$

$$= \left\{ \frac{\eta}{1-\eta} \cdot \frac{(\Sigma^* X_{i1}^2)(\Sigma^* X_{i2}^2) - (\Sigma^* X_{i1}X_{i2})^2}{\Sigma^* X_{i1}^2 + \Sigma^* X_{i2}^2 - 2\Sigma^* X_{i1}X_{i2}} \right\}^{1/2}$$

where Σ^* indicates that the summation is for all $i \neq 1$. Note in particular that as $X_{1j} \to \infty$ with $X_{11}/X_{12} = 1$, $\psi_\Omega(X_{1j})$ is bounded. Thus M-estimation with this resistant weighting cannot be dominated by the outlier-induced multicollinearity.

Another facet of outlier-induced multicollinearities is that multicollinearities can actually be strengthened when any of the resistant procedures suggested in the previous section are used. In fact, occasionally one induces a multicollinearity where none previously existed by weighting the predictor variable values. Extreme care must be exercised when these procedures are used; in particular, one should always examine the latent roots and latent vectors of the correlation matrix of predictor variables, variance inflation factors, etc. to determine whether multicollinearities have been induced or strengthened by the weighting of predictor variable values. The example discussed in the next section illustrates this point.

6. GASOLINE MILEAGE DATA

Hocking (1976), as part of an important and extensive survey of variabel selection techniques, utilized a set of data on gasoline consumption to illustrate the procedures he discussed. The original data set consists of a response variable, gasoline mileage (MPG) and ten predictor variables for each of 32 automobiles, the data taken from several issues of "Motor Trend" magazine. The ten predictor variables are engine shape (SHAPE), number of engine cylinders (CYL), automatic or manual transmission (AM), number of transmission speeds (GEAR), engine size (SIZE), engine horsepower (HP), number of carburetor barrels (CARB), final drive ratio (DRAT), weight (WT), and quarter mile time (TIME). Henderson and Velleman (1981) critize the use of MPG as the response variable, preferring to use $\mathrm{GPM} = (\mathrm{MPG})^{-1}$, and suggest that the preponderance of sports cars in the data base would make RATIO = HP/WT a potentially valuable addition to the set of predictor variables. After eliminating several predictor variables which do not appear to aid in the prediction of the response variable, we decided to illustrate the procedures discussed in the previous sections by regressing GPM on CYL, HP, DRAT, WT, AM, GEAR, and RATIO.

After examing various plots of the data to insure that no further transformations were apparent, several statistics were calculated for each observation as an aid to the detection of possible outliers. Table 1 displays these statistics along with a list of the automobiles included in the data set. The leverage values, eqn. (4.4), for the complete data set are shown in the second column of the table. The Lotus Europa, $h_{ii} = 0.872$, and the Maserati Bora, $h_{ii} = 0.681$, have the largest leverage values in the data set, both greatly exceeding Hoaglin and Welch's (1978) rough cutoff of $2(p+1)/n = 0.5$.

[Insert Table 1]

Also displayed in Table 1 are studentized deleted residuals, $t_{(-i)}$ (e.g., Gunst and Mason 1980, Section 7.1.3). These statistics measure the difference between an observed response Y_i and its predicted value $\hat{Y}_{(-i)}$ which is obtained using least squares coefficient estimates derived from the other (n-1) observations. Let SSE denote the residual sum of squares from the fit to GPM using all 32 observations. Then the ith studentized deleted residual is calculable as

$$t_{(-i)} = \frac{Y_i - \hat{Y}_{(-i)}}{\{\widehat{\mathrm{Var}}[\hat{Y}_{(-i)}]\}^{1/2}}$$

$$= \frac{(1 - h_{ii})^{-1/2} r_i}{\hat{\sigma}_{(-i)}}$$

where $(n-p-2)\hat{\sigma}^2_{(-i)} = SSE - (1-h_{ii})^{-1} r_i^2$. Individually the $t_{(-i)}$ follow a student t distribution with (n-p-2) degrees of freedom. Although the $t_{(-i)}$ are correlated, t tables can be used to furnish useful cutoff values for the detection of outliers.

The Chrysler Imperial has the largest studentized deleted residual in Table 1. Its value, $t_{(-i)} = -4.000$, places this statistic in the extreme lower tail of the corresponding t distribution and warrants a close examination of the Chrysler as a possible outlier. The Cadillac Fleetwood and Pontiac Firebird have moderately large studentized deleted residuals but are not so unusually large to be of concern in a sample of 32 observations.

Another statistic which will be examined as an aid in the deletion of outliers is Cook's (1977) distance measure. Let $\hat{\underline{\theta}}_{(-i)}$ denote the least squares estimator of $\underline{\theta}$ which is calculated from the (n-1) observations excluding the ith one. Cook (1977) defines a statistic

$$D_i = \frac{(\hat{\underline{\theta}} - \hat{\underline{\theta}}_{(-i)})' Z' Z (\hat{\underline{\theta}} - \hat{\underline{\theta}}_{(-i)})}{(p+1) MSE}$$

which allows a direct comparison of the least squares estimator from the complete data set, $\hat{\underline{\theta}}$, with the estimator calculated from (n-1) data points, $\hat{\underline{\theta}}_{(-i)}$. Although this statistic does not follow an F distribution, Cook suggested that F tables could still provide useful cutoff values for the detection of outliers. Because elimination of one observation from a homogeneous data set should leave $\hat{\underline{\theta}}_{(-i)}$ relatively unchanged from $\hat{\underline{\theta}}$, Cook further suggested that a value of D_i which is larger than a lower 10%. F value should be carefully studied as a possible outlier. We feel this criterion is often too conservative and choose to use a lower 25% cutoff value, $F_{.25}(8,24) = 0.623$. With this cutoff value the Lotus Europa is judged to have a strong influence on the estimation of $\underline{\theta}$. If the lower 10% F value, $F_{.10}(8,24) = 0.416$, is used the Chrysler Imperial would also be extremely influential on the estimation of $\underline{\theta}$.

The Cadillac, the Lincoln, and the Chrysler are the only American-made luxury cars which are included in this data set. They have very similar values on the predictor variables; e.g., they all have eight cylinder engines, they are the heaviest automobiles in the data set, etc. In fact, collectively they could be considered outliers because of their size relative to the other automobiles in Table 1. Yet individually their unusual features tend to be masked because they are similar among themselves; therefore, they do not induce individually large leverage values in Table 1. The Chrysler Imperial possesses a large studentized deleted residual and moderate-sized D_i because its gasoline mileage (hence, GPM) differs from the Cadillac and the Lincoln. The Chrysler's gasoline mileage is 14.7 while that of the Cadillac and Lincoln is 10.4 (see Henderson and Velleman 1981, Table 1). Since the Chrysler is an outlier due to an unusual response value, M-estimation should compensate for its influence on the fit.

The Lotus Europa poses different problems. Its leverage value suggests that the predictor variables for the Lotus are unusual and M-estimation alone might be unable to satisfactorily adjust for the ill effects of the Lotus. The Lotus is an outlier in predictor variable space because of the inclusion of RATIO as a predictor variable. The Lotus has relatively small values on HP and WT but, unlike other automobiles in Table 1 which also have small values on HP and WT, it possesses an unusually large value of RATIO. Other automobiles in Table 1 also possess large values on RATIO but they have large values on HP and WT as well. The Lotus is a three-dimensional outlier which the leverage values have aided in detecting.

Although the Maserati Bora also has a large leverage value, primarily due to its unusually large horsepower, it does not have correspondingly large values of $t_{(-i)}$ or D_i. This suggests that the Maserati is not unduly influencing the fit.

The last three columns of Table 1 display the leverage values, studentized deleted residuals, and Cook's distance values for a reduced data set in which the Chrysler and the Lotus are eliminated. Although the leverage value for the Maserati is now considerably larger than for the complete data set, the $t_{(-i)}$ and D_i values still do not indicate that the Maserati is unduly distorting the fit. Scanning the other $t_{(-i)}$ and D_i values in the last two columns does not lead one to conclude that any other observations in this data set are strongly influencing the fit.

In order to gauge the impact of the Chrysler and the Lotus on the coefficient estimates, least squares estimates for the complete data set are compared with those for the reduced (n=30) data set in the upper portion of Table 2. There are important differences in the significance (HP, AM, RATIO) and magnitudes (HP, DRAT, WT, AM, GEAR, RATIO) of the two sets of estimates. M-estimates, computed as described in Section 2 using initial least squares estimates of $\underline{\theta}$ and σ, are displayed in the lower portion of Table 2. The M-estimates for the complete data set and the reduced base set of observations are quite similar to the corresponding least squares estimates. This is reassuring for the base set but suggests that M-estimation for the complete data set has not successfully compensated for the inclusion of the two outliers. One would expect to see the M-estimates for the complete data set closer to the M-estimates for the base set than to the least squares estimates for the complete data set if M-estimation is adequately adjusting for these outliers.

[Insert Table 2]

The remaining columns of Table 2 exhibit least squares estimates and M-estimates for each of the predictor variable transformations which were discussed in Section 4. Overall, these "resistant" estimation schemes seem to perform worse than just using M-estimation on the (raw) complete data set predictor variables, with the possible exception of the weighted predictor variables in the last column. These estimates (obtained by setting $\eta = \bar{h} = .25$ for the Lotus) are quite similar in magnitude to the M-estimates for the complete data set but several of the coefficients are not significant when it appears they should be. Regardless of these comparisons, none of the resistant predictor variable transformations shows substantial improvement over M-estimation using the raw predictor variable.

The poor performance of the resistant estimators is attributable in part to a strong multicollinearity among the predictor variables. Inclusion of RATIO with HP and WT, while seemingly an important addition to the set of predictor variables, has induced a three-variable multicollinearity of the form

$$.53\ \text{RATIO} - .70\text{HP} + .45\ \text{WT} \approx 0.$$

This multicollinearity is detectable from the latent roots and latent vectors of $X_s'X_s$ and is further evidenced by the variance inflation factors of HP, WT, and RATIO: 51.6, 23.3, and 30.1, respectively. Note that the signs on the variables in the above multicollinearity are identical with the signs of the least squares estimates for the complete data set in Table 2. Elimination of the Chrysler and the Lotus worsens the problem since it strengthens the multicollinearity. The smallest latent root of $X_S'X_S$ drops form 0.0098 to 0.0024 for the base set and the three variance inflation factors increase to 217.0, 60.5 and 143.7, respectively. Note too that the magnitudes of the coefficient estimates for HP, WT, and RATIO all increase when the two outliers are removed from the complete data set and the signs of the coefficient estimates again correspond to those of the above multicollinearity. These sign patterns and large magnitudes are well-known characteristics of the ill effects of multicollinearities.

Each of the resistant estimators, despite their clear computational differences, either maintains or strengthens the multicollinearity among HP, WT, and RATIO. Due to the tendency for both outliers and multicollinearities to distort coefficient estimates, it would be fortuitous if any of the estimates in Table 2 were to accurately reflect the true relationship between GPM and these predictor variables. Since the multicollinearity is not outlier-induced, one cannot expect the resistant estimators to overcome the ill effects of the multicollinearity; indeed, several of the estimators in Table 2 exhibit the same sign pattern as the least squares estimates, although the magnitudes for the multicollinear predictor variables tend to be smaller than those given by the estimates for the base set.

Two changes were made in the data set in order to further examine these estimators. First, since the multicollinearity is due to a defined relationship between three predictor variables, viz, RATIO=HP/WT, one of these variables can be eliminated from the data set without seriously impairing prediction of GPM. RATIO was added to the data set because of the nature of the automobiles which are included in Table 1 and the belief that it might represent an important characteristic of the foreign sports cars. WT shows up as an important predictor variable in every analysis performed on these data. Consequently, we decided to eliminate HP and break up the induced multicollinearity. An alternative to this approach would be to retain all three predictor vairables and combine robust and biased estimation procedures (e.g., Askew and Montgomery 1980) but this alternative is beyond the scope of the present paper. The second change made in the data set was to increase the ratio variable on the Lotus Europa from 75 to 200 in order to accentuate the distortion it causes as an outlier.

With these two changes in the data set, HP removed and the Lotus' RATIO value set to 200, the outlier statistics for the complete and base 30 data sets are as shown in Table 3. The only large leverage value in the complete data set occurs for the Lotus Europa. The largest studentized deleted residual is associated with the Chrysler Imperial and the Lotus and the Chrysler both have D_i values which exceed a lower 25% cutoff value for an $F(7,25)$ distribution. In contrast, the outlier statistics for the base set reveal no strong indications of outliers.

[Insert Table 3]

Table 4 displays the least squares and the M-estimates for the same estimators as in Table 2. Again the least squares estimates and the M-estimates for the base set are quite similar. Unlike Table 2, the least squares estimates and the M-estimates are not virtually identical for the complete data set. The M-estimates for the complete data set are closer to the M-estimates for the base set than are the least squares estimates. With HP removed, the M-estimates do appear to be reducing the effect of the large residuals. The two observations which have $\phi(r_i/\hat{\sigma})$ values less than 1.0 in the last iteration of eqn (2.12) correspond to the Chrysler Imperial and the Pontiac Firebird, the two observations which have the largest $t_{(-i)}$ values in the third column of Table 3. Interestingly, the Lotus Europa is not weighted by M-estimation on the complete data set. The effect of the Lotus on the coefficient estimates is unaltered by

M-estimation.

[Insert Table 4]

The remaining estimators in Table 4 are the same as those in Table 2 except for the weighted estimates using $\psi_{\Omega}(\cdot)$. In studying the M-estimates for all the estimators in Tables 2 and 4, it appeared that the M-estimates either would fail to weight or would not adequately weight residuals whose leverage values were not sufficiently small, even if the residual appeared to be large. In other words, if a residual was large it would only get weighted by M-estimation if its leverage value was suitably small. Adequate weighting by M-estimation seemed to require that the leverage value be no larger than the average of all the leverage values, $\bar{h} = (p+1)/n$. In applying the weighted estimator (4.7) we decided to weight the rows of X corresponding to both the Chrysler Imperial and the Lotus Europa so their leverage values would equal 0.20 ($\bar{h} = 0.22$).

The first three "resistant" estimators shown in Table 4 still fail to improve on the M-estimates for the raw predictor variables in the complete data set. Each of these estimates attempts to adjust for outliers by modifying the observations on a single predictor variable or a single principal component without regard to the values on the other predictor variables or prinicpal components. In each case some of the coefficient estimates appear to be close to those of the base set but others are not. Since the predictor variables are not orthogonal, robust regression procedures which weight variables or components individually might not only be incapable of adequately adjusting for outliers but they could also further distort the estimates by changing the correlation structure among the predictor variables. These first three estimators could be suffering such a problem.

The last estimator does seem to adequately adjust for the outliers in this data base. After the predictor variable values for the Chrysler and the Lotus are weighted so their leverage values are approximately 0.20, the M-estimates weight the residuals corresponding to both of these observations and the Pontiac Firebird. The resulting coefficient estimates are the most similar to the base set in Table 4.

7. CONCLUDING REMARKS

The need for developing regression procedures which are both robust to error assumption violations and resistant to aberrant predictor variable values has been demonstrated in the theoretical derivations of Sections 2 and 3 and the examples discussed in the previous section. Further studies are needed before any of the procedures discussed in this paper can be recommended for general use but the example suggests several important properties which good resistant estimators should possess. First, they must be able to adjust for outliers in the predictor variables without substantially altering the correlation structure which is imposed by the "non-outlier" observations. Whether weighting schemes which operate on columns of X rather than its rows can accomplish such an adjustment without altering the true underlying correlation structure remains to be carefully investigated. A second property of robust/resistant estimators which seems desirable is that the estimators should be capable of weighting large residuals even if in the raw data set the residuals are accompanied by large leverage values. If M-estimation is used on the residuals, this might require a weighting of observations to insure that leverage values are sufficiently small. Small leverage values are not only desirable for accurate estimation but also, as Huber (1981, Chapter 7) proves, required for asymptotic normality of the estimators for nonnormal errors.

Rank transforms offer another possible alternative to the procedures studied in this article. Rank transforms have been shown by Iman and Conover (1979) to be effective robust alternatives for prediction of the response variable but not necessarily for parameter estimation, the focus of this paper.

ACKNOWLEDGEMENT

The research reported in this article was supported by the Office of Naval Research Contract No. N00014-75-C-0439.

TABLE 1. OUTLIER STATISTICS FOR GASOLINE MILEAGE DATA. (Prediction of GPM)

Automobile Type	Complete Data Set			Base Set (n = 30)		
	h_{ii}	$t_{(-i)}$	D_i	h_{ii}	$t_{(-i)}$	D_i
Mazda RX-4	.166	.162	.001	.177	.066	.000
Mazda RX-4 Wagon	.184	-.258	.002	.187	-.523	.008
Datsun 710	.168	.534	.007	.202	.379	.005
Hornet 4 Drive	.119	-.677	.008	.125	-.845	.013
Hornet Sportabout	.126	-.994	.018	.134	-1.231	.029
Valiant	.224	.377	.005	.228	.508	.010
Duster 360	.255	.614	.017	.258	.611	.017
Mercedes 240D	.316	-.329	.007	.346	-.219	.003
Mercedes 230	.276	-.341	.006	.278	-.651	.021
Mercedes 280	.237	-.107	.000	.238	-.239	.002
Mercedes 280C	.237	.559	.012	.238	.610	.015
Mercedes 450SE	.080	-.935	.009	.092	-1.453	.026
Mercedes 450SL	.092	-.805	.008	.094	-1.027	.014
Mercedes 450SLC	.088	.304	.001	.091	.373	.002
Cadillac Fleetwood	.265	2.342	.208	.358	1.485	.146
Lincoln Continental	.318	1.802	.173	.443	.625	.040
Chrysler Imperial	.298	-4.000	.522			
Fiat 128	.164	-.809	.016	.243	-.705	.020
Honda Civic	.381	.413	.014	.423	.927	.079
Toyota Corolla	.140	-.500	.005	.145	-.380	.003
Toyota Corona	.382	.729	.042	.519	.477	.032
Dodge Challenger	.215	1.109	.042	.222	1.933	.119
AMC Javelin	.162	1.230	.036	.165	1.798	.073
Camaro Z-28	.328	.839	.044	.360	.588	.025
Pontiac Firebird	.088	-1.938	.041	.092	-2.747	.073
Fiat X1-9	.149	.505	.006	.167	1.033	.027
Porsche 914-2	.170	.474	.006	.193	.623	.012
Lotus Europa	.872	-1.394	1.590			
Ford Pantera L	.364	.120	.001	.480	-.532	.034
Ferrari Dino 1973	.236	.288	.003	.425	.128	.002
Maserati Bora	.681	-.227	.014	.806	.307	.051
Volvo 142E	.217	-.189	.001	.271	-1.195	.065

TABLE 2. COMPARISON OF GPM COEFFICIENT ESTIMATES ($x10^{-3}$).

Least Squares Estimates

Predictor Variable	Base Set (n = 30)	Complete Data Set	Huber's $\psi^j(\cdot)$	Mallows $\psi_M(\cdot)$	Principal Component	Weighted $\psi_\Omega(\cdot)$
CYL	-.958	.459	-.673	4.181*	.571	.513
HP	-.309*	-.071	.242*	-.193	-.202	-.046
DRAT	7.042*	3.389*	-3.222	2.799	2.083	3.285
WT	31.664*	17.747*	6.768	14.992	24.000*	16.766*
AM	7.553*	4.468	4.307	3.009	5.767	4.450
GEAR	-8.966*	-5.118*	-2.280	-7.505	-2.225	-4.936
RATIO	1.378*	.491	-.672*	.127	1.037	.402

M-Estimates

Predictor Variable	Base Set (n = 30)	Complete Data Set	Huber's $\psi^j(\cdot)$	Mallows $\psi_M(\cdot)$	Principal Component	Weighted $\psi_\Omega(\cdot)$
CYL	-.782	.078	-.213	4.181*	.148	-.013
HP	-.303*	-.096	.226*	-.193	-.234	-.146
DRAT	6.960*	4.794	-2.235	2.799	3.838	4.804
WT	31.278*	20.855*	5.607	14.992	27.280*	22.543*
AM	7.523*	5.941*	2.710	3.009	7.165	5.747
GEAR	-8.937*	-6.371*	-2.147	-7.505	-3.736	-6.564*
RATIO	1.348*	.588*	-.582	.127	1.178*	.769

*Significant at an $\alpha = .20$ (two-tailed) level.

TABLE 3. OUTLIER STATISTICS FOR ALTERED GASOLINE MILEAGE DATA. (Prediction of GPM)

Automobile Type	Complete Data Set			Base Set (n = 30)		
	h_{ii}	$t_{(-i)}$	D_i	h_{ii}	$t_{(-i)}$	D_i
Mazda RX-4	.120	-.055	.000	.154	.289	.002
Mazda RX-4 Wagon	.123	-.540	.006	.185	-.447	.007
Datsun 710	.156	.725	.014	.186	.559	.011
Hornet 4 Drive	.120	-.629	.008	.124	-.772	.012
Hornet Sportabout	.125	-.967	.019	.127	-1.067	.024
Valiant	.226	.318	.004	.228	.496	.011
Duster 360	.127	1.018	.021	.258	.564	.016
Mercedes 240D	.288	-.568	.019	.296	-.578	.021
Mercedes 230	.262	-.229	.003	.276	-.723	.029
Mercedes 280	.209	-.327	.004	.238	-.205	.002
Mercedes 280C	.209	.305	.004	.238	.626	.018
Mercedes 450SE	.073	-1.025	.012	.089	-1.507	.030
Mercedes 450SL	.091	-.841	.010	.092	-1.063	.016
Mercedes 450SLC	.086	.224	.001	.089	.294	.001
Cadillac Fleetwood	.245	2.150	.187	.353	1.581	.183
Lincoln Continental	.303	1.707	.168	.440	.706	.057
Chrysler Imperial	.313	-3.776	.606			
Fiat 128	.134	-.977	.021	.138	-1.149	.030
Honda Civic	.375	.177	.003	.399	.602	.035
Toyota Corolla	.136	-.530	.007	.139	-.482	.006
Toyota Corona	.226	1.337	.072	.432	.983	.105
Dodge Challenger	.194	.795	.022	.220	1.782	.117
AMC Javelin	.132	.957	.020	.160	1.873	.086
Camaro Z-28	.276	1.178	.075	.348	.375	.011
Pontiac Firebird	.082	-1.996	.045	.090	-2.724	.082
Fiat X1-9	.134	.393	.004	.136	.717	.012
Porshe 914-2	.177	.384	.005	.180	.777	.019
Lotus Europa	.921	-1.520	3.666			
Ford Pantera L	.350	.327	.009	.359	.127	.001
Ferrari Dino 1973	.278	.413	.010	.301	.696	.030
Maserati Bora	.315	.168	.002	.484	-.912	.112
Volvo 142E	.194	.090	.000	.241	-.854	.034

TABLE 4. COMPARISON OF GPM COEFFICIENT ESTIMATES ($x10^3$), ALTERED DATA BASE.

Least Squares Estimates

Predictor Variable	Base Set (n = 30)	Complete Data Set	Huber's $\psi^j(\cdot)$	Mallows $\psi_M(\cdot)$	Principal Component	Weighted $\psi_\Omega(\cdot)$
CYL	-.757	2.251*	.072	4.478*	2.307	1.653
DRAT	4.865*	2.917	.917	1.498	1.339	4.449
WT	19.312*	13.895*	22.114*	13.048*	16.052*	16.407*
AM	7.053*	4.825	8.847*	6.504	5.447	6.422
GEAR	-6.580*	-2.478	-3.043	.003	1.578	-3.804
RATIO	.303*	.067*	.277	-.150*	.103	.093

M-Estimates

Predictor Variable	Base Set (n = 30)	Complete Data Set	Huber's $\psi^j(\cdot)$	Mallows $\psi_M(\cdot)$	Principal Component	Weighted $\psi_\Omega(\cdot)$
CYL	-.584	1.865*	.888	4.342*	2.155*	.807
DRAT	4.825*	4.199	1.548	2.067	3.006	4.106
WT	19.164*	15.944*	18.151*	14.083*	17.918*	16.853*
AM	7.032*	6.180*	5.696	7.318	6.216	6.169*
GEAR	-6.598*	-3.645	-2.516	-.366	.881	-4.680*
RATIO	.295*	.083*	.299	-.147*	.109*	.179*

*Significant at an $\alpha = .20$ (two-tailed) level.

REFERENCES

Andrews, D. F. (1974), "A Robust Method for Multiple Linear Regression," Technometrics, 16, 523-531.

Andrews, D. F., et. al., (1972). Robust Estimates of Location: Survey and Advances. Princeton University Press, Princeton, New Jersey.

Askin, R. G., and Montgomery, D. C. (1980), "Augmented Robust Estimators," Technometrics, 22, 333-341.

Box, E. P. and Watson, G. S. (1962), "Robustness to Non-Normality of Regression Tests," Biometrika, 49, 93-106.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," Technometrics, 19, 15-18.

Denby, L. and Larsen, W. A. (1977) "Robust Regression Estimators Compared via Monte Carlo," Comm. Statist., A6, 355-362.

Dutter, R. (1975), "Robust Regression: Different Approaches to Numerical Solutions and Algorithms," Res. Rep. No. 6, Fackgruppe für Statistik, Eidgen., Technische Hochschule, Zurich.

Dutter, R. (1977), "Numerical Solution of Robust Regression Problems: Computational Aspects, a Comparison," J. Statist. Comp. and Simul., 5, 207-238.

Gunst, R. F. and Mason, R. L. (1980). Regression Analysis and its Application. New York: Marcel Dekker, Inc.

Hampel, F. R. (1968), "Contributions to the Theory of Robust Estimation," Ph.D. Thesis, University of California, Berekley.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," J. Amer. Statist. Assoc., 69, 383-393.

Henderson, H. V. and Velleman, P. G. (1981), "Building Multiple Regression Models Iteratively," Biometrics, 37, 391-411.

Hoaglin, D. C. and Welsch, R. E. (1978), "The Hat Matrix in Regression and ANOVA," The American Statistician, 32, 17-22.

Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," Biometrics, 32, 1-44.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," Ann. Math. Statist., 35, 73-101.

Huber, P. J. (1981). Robust Statistics. New York: John Wiley and Sons, Inc.

Iman, R. L. and Conover, W. J. (1979), "The Use of the Rank Transform in Regression," Technometrics, 21, 499-509.

Koenker, R. and Bassett, G., Jr. (1978), "Regression Quantiles," Econometrica, 46, 33-50.

Mallows, C. L. (1973), Talk presented at Conference on Robust Regression held at NBER, Cambridge, Massachusetts.

Rupert, D. and Carroll, R. J. (1980), "Trimmed Least Squares Estimation in the Linear Model," J. Amer. Statist. Assoc., 75, 828-838.

Seber, G. A. F. (1977). Linear Regression Analysis. New York: John Wiley.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 157

2. GOVT ACCESSION NO.: AD-A117022

3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): Robust Regression Procedures for Predictor Variable Outliers

5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL REPORT

6. PERFORMING ORG. REPORT NUMBER: 157

7. AUTHOR(s): Dovalee Dorsett and Richard F. Gunst

8. CONTRACT OR GRANT NUMBER(s): N00014-82-K-0207

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Southern Methodist University
Dallas, Texas 75275

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 042 479

11. CONTROLLING OFFICE NAME AND ADDRESS:
Office of Naval Research
Arlington, Va. 22217

12. REPORT DATE: March 1982

13. NUMBER OF PAGES: 44

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):

15. SECURITY CLASS. (of this report):

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report):

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purposes of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):

18. SUPPLEMENTARY NOTES:

19. KEY WORDS (Continue on reverse side if necessary and identify by block number):

M-estimation, Resistant Estimators, Multicollinearity

20. Abstract. Least squares estimators of regression coefficients can be overly sensitive to violations of certain errors assumptions; e.g., outliers in the response variable. One solution to the presence of outliers in a data base is to apply univariate robust estimation procedures to the residuals of estimated models. Equally problemmatic as outliers among the response variable are outliers or aberrant values for the predictor variables or an unusual combination of predictor variable values for a few observational units can distort least squares estimators even if the error assumptions are valid. This article discusses robust regression procedures, with special emphasis on techniques which are resistant to extreme predictor variable values.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)